The Lessac Technologies Hybrid Concatenated System for Blizzard Challenge 2013

Reiner Wilhelms-Tricarico, John Reichenbach, Gary Marple

Lessac Technologies, Inc., USA

{reiner.wilhelms, john.reichenbach, gary.marple} @lessactech.com

Abstract

Lessac Technologies has developed a technology for concatenated speech synthesis based on a novel approach for describing speech in which expressivity, voice quality, and speaking style are fundamental. The main aspect of our system is that instead of traditional phonetic symbols, we use a much more fine-grained and richer set of entities called Lessemes to describe speech and to label units, which allow a richer and more precise characterization of speech sounds. The front-end portion of our synthesizer translates plain input text into a sequence of these units by syntactic parsing and applying a set of rules developed from expertise. We use a Bayesian method to obtain a particular trainable mapping from linguistic and prosodic features encoded in the Lessemes to a trajectory in the acoustic parameter space. Unit selection consists of selecting the best candidate units from a data base to match them to the target trajectory, while minimizing discontinuities between them. For the voice used in this Challenge we implemented a preliminary model for remapping intonation features called levels based on acoustic features.

Index Terms: Speech Synthesis, Blizzard Challenge, Lesseme.

1. Introduction

This is our 4th entry to the Blizzard Challenge. For 2013 the voice recordings that served as a basis for the Challenge were supplied by Lessac. The voice recordings were audio books narrated by a speaker named Cathy. About 20 hours of voice data were supplied in wave file format with associated text segmented to the sentence level; this is the same data that Lessac had previously used to make our Cathy voice. Lessac supplied over 175 hours of additional voice recordings narrated by the same speaker in mp3 format without associated text. The competition consisted of two parts. In the first task, EH1, each competitor built a voice making use of only the voice material that we had used to make our first voice of Cathy; therefore we simply used without modification our existing Cathy voice.

For the second task, EH2,
we extended the voice by 38 percent by adding the material from 4825 additional prompts to the 9728 prompts included in our original voice, making it 14553 prompts in total. These prompts are usually individual sentences but often the prompts can be two or more shorter sentences. For the most part, we used the same technology for this extended voice, but we made some effort to adjust the annotations of both the old prompts and the new prompts by mainly statistical methods, described in detail in this report.

2. Approach to Corpus Selection

Lessac has built a half dozen TTS voices from similar audio book data. We use a semi-automated tool we have developed to segment the voice recordings into sentences or prompts. The user listens to the voice recordings while simultaneously being presented with the text, on a sentence by sentence basis, and then repeatedly presses a key as he hears the end of each sentence. Since this cannot be done entirely accurately the first time through, this is followed with a second clean-up tool where the user is presented with a small amount of wave data (both graphically and audibly), and he can add, move or delete each sentence segmentation break. However, in order to get clean data, it requires listening to the equivalent of the entire recording at least twice. We expanded the original database provided to all participants by adding the entire three part series from Pride and Prejudice, from which a total of 4825 prompts were used.

3. Lessac Technologies Text-to-Speech System

Similar to other systems, the Lessac Technologies text-to-speech system consists of two main components: the front-end, which takes plain text as input, and outputs a sequence of graphic symbols, and the back-end, which takes the graphic symbols as input to produce synthesized speech as output. In what follows, we briefly discuss the properties that distinguish our system from others and, we believe, play an important role in producing expressive synthesized speech.

3.1 Use of Lessemes

Successful production of natural sounding synthesized speech requires developing a sufficiently accurate symbolic set of sound representations that can be derived from the input text, and that relate the input text to be pronounced with the corresponding synthesized speech utterances that are heard by the listener. Rather than adopting traditional symbolic representations, such as IPA, SAMPA, or ARPAbet, Lessac Technologies has derived an extended set of symbolic representations called Lessemes from the phonosensory symbol set for expressive speech as conceived by Arthur Lessac [1]. The Lesseme system for annotating text explicitly captures the musicality of speech, and from the start avoids the artificial separation of prosodic and linguistic features of speech.

In their basic form and meaning, Lessemes are symbolic representations that carry in their base form segmental information just like traditional symbolic representations. To be able to describe speech more accurately and to include in the symbol set information that is not carried by a typical phonetic symbol, each base Lesseme can be sub-typed into several more specific symbols which then represent phonetic information found in traditional phonetic symbols plus descriptors for coarticulation and suprasegmental information. Acoustic data demonstrate different properties of a set of Lessemes which are normally collapsed under one phonetic label in other systems [2].

For General American English, with the present Lesseme specification, there can be as many as 1,500 different Lessemes. Compared to other sets of representations which usually contain about 50 symbols, Lessemes allow more fine-grained distinction of sounds. Units of the same type share closely similar acoustic properties. By having supra-segmental information directly encoded in Lessemes, we believe our system can target available units for concatenation better than a system with a relatively impoverished intonation annotation scheme. This should be useful especially when trying to produce expressive speech from a very large database.

3.2. Front-end with extensive linguistic knowledge

The front-end which derives Lessemes from plain text input is a rules-based system. The rules are based on expert linguistic knowledge from a wide variety of fields including phonetics, phonology, morphology, syntax, light semantics, and discourse. Simplistically, the Lessac front-end labels text, building from, at the lowest level, letters, spaces and punctuation marks. These letters, spaces and punctuation marks are interpreted by the front-end, and assembled as syllables, words, phrases, sentences, and paragraphs to be spoken, along with context-aware labeling for appropriate co-articulations, intonation, inflection, and prosodic breaks.

First, the input text is processed by a syntactic parser which generates the most likely syntactic tree for each sentence, and tags words with part-of-speech (POS) information. In the next step, words are transcribed by use of a pronunciation dictionary into base Lessemes accompanied by lexical stress. Homograph disambiguation based on POS tags takes place at this step. Subsequent processing steps modify the base Lessemes by making successive decisions based on the overall phrase and sentence structure. In particular, prosodic breaks are inserted in meaningful places by taking into consideration factors such as punctuation, phrase length, syntactic constituency, and balance. In most phrases, an operative word is marked which carries the highest pitch prominence within the phrase. In addition, Lessemes are assigned inflection profiles and one of two degrees of emphasis. Context-based co-articulations across word boundaries are also captured. The result is a full Lesseme for each sound which encodes expressive intonational content in addition to segmental information found in traditional phonetic symbols.

The front-end process is able to develop a complete Lesseme label stream with plain normally punctuated text as the sole input. This Lesseme stream is delivered to the signal processing back-end.

3.3. Voice database construction

In addition to the machine readable form used as the input to the signal processing back-end, Lessemes are also used in creating new voices, namely to automatically generate a human readable graphic output stream which can be thought of as annotated text plus a musical score, as illustrated in figure 1.

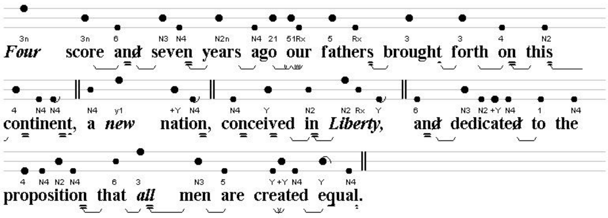
3n 3n 6 N3 N4 N2n N4 21 51Rx 5 Rx 3 3 4 N2
Four score and seven years ago our fathers brought forth on this
4 N4 N4 N4 y1 +Y N4 N4 Y N2 N2 Rx Y 6 N3 N2 +Y N4 1 N4
continent, a new nation, conceived in Liberty, and dedicated to the
4 N4 N2 N4 6 3 N3 5 Y +Y N4 Y N4
proposition that all men are created equal.

Figure 1: *Lessac Technologies annotated text*

In the annotation, vowel orthographic forms are designated with Arthur Lessac's phonosensory symbols. Consonant orthographic forms are marked with information indicating whether the consonant is sustainable (double underlined) or percussive, i.e. pronounced with a brief contact within the mouth (single underlined), as well as how the consonant is linked to the next sound in connected speech. The musical score on top of the orthographic forms depicts notes which represent the intonation pattern that a person with sufficient voice training can follow. Each syllable corresponds to a note. Higher notes are pronounced with higher pitch. Large notes define stressed syllables while small notes refer to unstressed syllables. Some notes are further specified with an inflection, which reflects a particular shape of pitch movement within the syllable.

During the voice database construction, the text to-be-recorded is first processed by the front-end, yielding the stream of Lessemes. In building a full Lessac voice, the resulting stream is then transformed into a human readable form, as seen in figure 1, which we use as the combined script and score for the trained voice talent during the recordings. The way the voice talent records the prompts is controlled by the annotated text and musical score. The recordings of the prompts are then segmented and labeled with the same Lessemes that underlie the script and score that the voice talent followed. The fact that the same Lessemes are output for the voice talent script as well as the labeling of the database creates a direct link between each speech snippet and its Lesseme label, thus a high degree of correspondence between the symbols and the sounds as actually recorded by the voice talent.

However, for Lessac TTS voices constructed from audio book data, such as the "Greenman" voice from last year's challenge and even more so for the current "Cathy" voice, such a high degree of symbol-to-sound correspondence is not guaranteed. Using our current voice-building techniques, unless the voice building process includes a labor intensive manual notation process, the symbol to sound correspondence reflects only the expert knowledge contained in our front-end. Our front-end prosody represents an idealized "reportorial" prosody, which although relatively accurate for most speakers, is only one of many speaking styles that a voice actor could use to read the text.

We make use of this correspondence in the unit selection process by evaluating units in the data base according to the context dependent linguistic and prosodic features, in order to preselect a subset of unit candidates, which are then evaluated by the model described in the following.

3.4. Hierarchical Mixture of Experts for mapping linguistic features to acoustic parameters

To enhance methods for target cost calculation and unit selection, we apply the Hierarchical Mixture of Experts (HME) model [3] [4] to learn the parameters of a statistical model of the relationship between the Lesseme representation of the input text and the ideal acoustic observables in the recordings.

A functional diagram of the HME model is shown in figure 2.

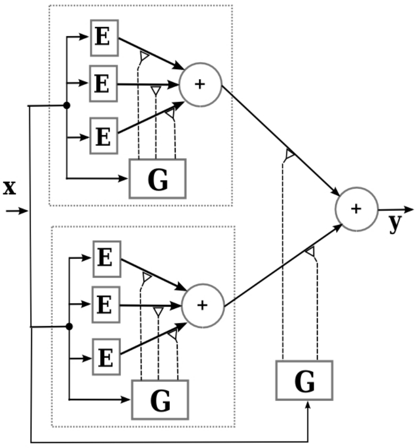

Figure 2: *Hierarchical Mixture of Experts model. (E: experts, G: gates, x: input, y: output)*

The HME model applied to the problem of mapping prosodic features to acoustic observables makes use of the interpretation of the model as a parameterized mixture of Gaussians. Each expert in the model represents one multi-dimensional normal distribution with a variable expectation vector that depends on the input x. The parameters for each expert also include a full covariance matrix that is estimated and updated during the training process. Each block of experts in a group or clique (Figure 2 shows 3 experts in each of 2 cliques) together with a gating network represent one mixture of Gaussians whereby the mixture coefficients are computed in the gates as a function of the input. Multiple groups of experts can be combined by another gate in a similar way. The complete network represents a mixture of Gaussians whose parameters are trained from pairs of known input and output. During the learning process, the parameters in the experts and gates are adjusted so that, for a given known input x, the probability of obtaining the desired known output y is maximized over all available data.

In our application of the HME model, the input x includes the linguistic and prosodic features and the output y are acoustic observables, which include MFCC's, F0, duration, and intensity, mostly the same type of parameters used in database segmentation, see 4.3 below. The model is applied and trained as a recurrent system, which means that the predictions of acoustic observables, y[n], for one sound at time index n are included in the input x[n+1] for the prediction of the next y[n+1].

We use supervised learning with the HME model to map linguistic feature sequences to a trajectory in the acoustic parameter space, which is represented by via points and for some of the parameters their velocity or rate of change. The structure of the model is shown in figure 3. The system steps through a sequence of Lessemes and predicts for each Lesseme the vector of acoustic parameters that specify the unit, whereby the input to the model consists of the feature information of the previous, the current and the next two Lessemes. Further, by feeding back the previously predicted acoustic parameter vectors as input to the model, the model becomes partially auto-regressive. This facilitates the learning task because the model only has to learn to predict the current acoustic vector conditioned on the last two acoustic vectors and the input linguistic features. Learning proceeds in two phases. Initially, the looped-back input to the model consists of the actual acoustic vectors until the model begins to converge. Then, training is continued by having the predictions for the last two time slots become inputs for the prediction of the current time slot. Learning then proceeds by repeatedly processing a large number of sentences in the database, until the error variance cannot be lowered further.

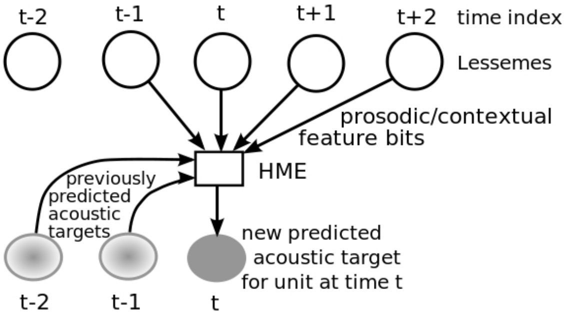
t-2
t-1
t
t+1
t+2
time index
Lessemes
prosodic/contextual feature bits
HME
previously predicted acoustic targets
new predicted acoustic target for unit at time t
t-2
t-1
t

Figure 3: *Recurrent and partially auto-regressive prediction of intonation contour and other acoustic targets by HME*

During the target cost calculation process, we compute the cost as the distance of the acoustic parameters of a candidate unit from the ideal trajectory, which is in turn directly predicted from the linguistic feature variables. This distance measure makes use of the predicted mixture covariance matrix which is obtained by combining the experts' covariances according to the gating weights, (see Figure 2). To reduce processing time, we reduce the number of candidates first by applying a rapid search with binary patterns generated from some of the features, and then compute the exact target cost for a smaller subset of close candidates. Since the HME provides the parameters of a probability density in the acoustic parameter space, we compute for the remaining candidates their probability under this distribution and use as target cost a penalty that is proportional to the negative logarithm of the candidates' probability.

Using the Lesseme representation of speech sounds, the output of the front-end results in a large number of features, which is augmented further by bundling neighboring features as shown in figure 3. The HME model overcomes the sparsity problem in the acoustic database by mapping the Lesseme features and context onto the acoustic parameter space as a target trajectory. At the same time it automatically provides a variable metric near the target trajectory, against which the candidates in the data base are matched during unit-selection.

4. Building the 'Cathy' Voice

In the following we describe the work done on extending our existing 'Cathy' voice to create the voice used for the task EH2; for task EH1 we used our existing Cathy voice without any changes. Most of the techniques we used here are the same as those already applied in previous years, in particular a new method of pitch marking that we developed last year, and described in more detail in last year's report, was used with little change.

4.1. Automatic Relabeling of Lessemes.

The prosody of the audio recordings by Cathy are very often in contradiction to the prosodic rules that we are using in our frontend. The type of prosody that we call "reportorial" is a standard we had previously developed and used in all other voices. The core issue here is that our rule based prosody is supposed to be used in annotating the prompts before they are recorded. According to our design, the voice model used for recording should be trained in reading the specific annotation which is automatically generated by the frontend, as described in the previous section. So far, we have only done this our first voice, Nancy. For Nancy, the prompts were recorded after rehearsing and careful checking of inconsistencies between prosodic annotation and meaning. In some cases, the annotations were then corrected if the rule system used in the frontend resulted in phrasing or stress patterns that were in contradiction to the intention and meaning in a sentence as perceived by the speaker. This was not the case in last year's competition with Greenman's Librevox voice, but our results were still quite good (2nd place in the overall ranking).

By studying the recording material we found that the discrepancies between the frontend's output annotation of the material and the actually produced speech was much more pronounced for Cathy's recordings then it was for Greenman's voice. To overcome this issue, we had to choose between a number of options. We could either design a new rule system that would be more appropriate for Cathy's recordings, or we had to select only recorded sentences that followed our prosody rules, usually just narrative passages. A third option was to try to relabel many of the Lessemes so that their physical acoustic characteristics were more properly represented in the labels. The first option was too difficult to realize, given our resources and the time allocated to this task: The readings by Cathy are extremely variable. There are many long narrative passages that could fit within our already existing prosodic model labeled "reportorial". But there are also many passages that can not really be subsumed under a single prosodic model. In particular, the audio books contain many dialogue situations where Cathy reads the text in a highly expressive manner, often emulating or mimicking some quite eccentric people and facilitating the listener's differentiation of their different idiosyncrasies. Given the variety in the recordings, we would need to build not only one new prosodic model but several. The second option, namely selecting only passages that we considered to be within the reportorial prosody was not successfully pursued because it would have required many hours of listening, re-listening, and reading and re-reading of the text. The main obstacle here was that there were simply too many passages where it was too difficult to make consistent decisions. The third option, automatic relabeling, is an experimental procedure requiring the least amount of direct listing and editing, so we built a preliminary implementation of this as the basis of our EH2 voice first.

Methods.

The assumption is made that (binary) linguistic features obtained from the analysis of the text and used in the synthesizer for unit selection are more or less directly correlated with physical features such as intensity and pitch as well as spectral features and the dynamics of their changes. The correlation of individual features with physical acoustic parameters can be expected to be fairly weak, with a few exceptions. Yet, in combination, there is mutual information between the linguistic features and the physical acoustic observables. Making this assumption we attempted to "fix" some of the linguistic features for the Cathy voice. A particularly important feature in our synthesizer is the level feature; levels are mainly understood as grades of prominence. In the current frontend rules for reportorial prosody, the level feature has only three values, {1,2,3}. They are effectively applied to a syllable but are attached to vowels in the syllable nucleus. A separate feature is stress which can have the values {L,H} and corresponds roughly to lexical stress. The level values are distinct from lexical stress but there is interaction between lexical stress and the level that a vowel receives in context, but the lexical stress is usually not variant (it is directly derived from the dictionary together with, in some cases, part of speech information). The connection between acoustic features and the level feature is not straight forward and highly context dependent, and it can not simply be derived from just taking F0 and intensity (energy) into account.

Since Cathy's voice was not recorded under the rule system that we used for our first voice, Nancy, it cannot be expected that the level feature (and several others) can be properly assigned by the frontend. So we built a simple information structure that relates some of the linguistic features with acoustic features and with themselves, and trained the details on the date from Nancy's voice which is by definition properly annotated. The hypothesis was that by using some of the linguistic features and some of the acoustical features, other linguistic features could be predicted. This model was realized by a combination of an auto-encoder or auto-associator and a multi-layer perceptron, see figure 4. The auto-associator was trained in unsupervised training to represent the high dimensional and sparse linguistic feature vectors in a lower dimensional space, without significant loss of information, see [5]. We used several of the intrinsic features of the Lessemes of three consecutive segments, and some more supra-segmental features. The multi-layer perceptron has two types of input, the output from the auto-encoder, and some acoustic features measured for three consecutive segments, namely the current, the preceding and the following segment. It was trained to predict merely the level of the central Lesseme. The following linguistic features were used for consecutive Lessemes: For vowels: (vowel type {short, long, diphthong, schwa}, vowel height; vowel frontness; lip rounding; vowel lexical stress). For consonants (12 bit consonant type pattern; a place feature;

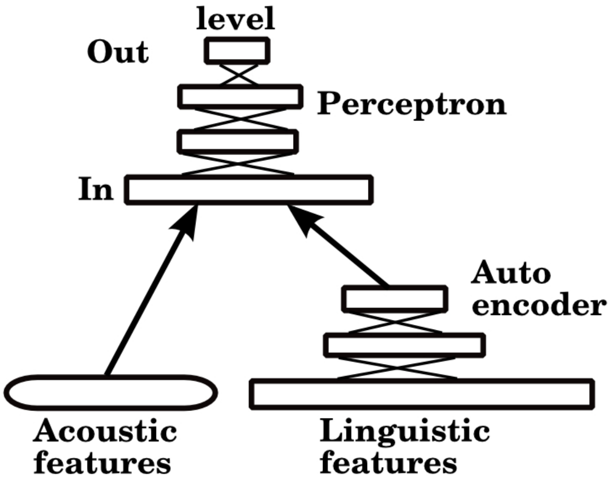

Figure 4. *Network structure for computing level information. The input of the perceptron is a combination of acoustic features and the dimensionally reduced linguistic features.*

consonant sonority; voiced/unvoiced information). Pauses were also represented, using a choice of types represented by binary features. In these input features we attempted to use only features that do not depend on the level features. However, we used the lexical stress feature which is assumed here to be invariant from the prosody - as we are using the same dictionary for Cathy as for Nancy. However there are inherent correlations between the linguistic features, so the possibility cannot be completely excluded that the neural network learns to some extent rules that are imposed by the frontend.

The acoustic input variables for the network were derived from the measured durations, F0, and log intensity contours. Both intensity and F0 are scaled in such a way that the first two moments of their distributions are matched between the two voices. This is a linear transformation between the two found by linear fitting cumulative histograms, with the result that for example an F0 value for Nancy of 200 Hz corresponds to 186Hz, and the axes are scaled relative to each other by a scale of 0.962. The perceptron has three outputs, which are obtained by applying a sigmoid function to the weighted sums of their inputs. These three outputs are understood as the likelihood of the 1st, 2nd or 3rd level.

The two networks were separately trained. The auto-encoder is trained based on unsupervised learning that is related to the training of reduced Boltzmann machines, and is as described in reference 5. After the learning of the auto-encoders is completed, its parameters are kept fixed, and the perceptron which receives the output of the auto-encoder and the acoustic parameters as input is trained by back propagation.

To make it less likely that consecutive Lessemes could receive both level 3 or that there were too many jumps between levels, the output of the perceptron was used as input to a Viterbi algorithm that determines the final level of each vowel. For this, the parameters of the simplest possible Markov model for level information were trained from Nancy's data. The Markov model had only three states corresponding to the levels 1, 2 or 3. The data from Nancy's voice were used to compute a matrix of transitional probabilities between the levels in subsequent vowels, ignoring any further contextual information. A Viterbi algorithm then combines the level probability from the above explained network structure and the transitional probabilities, and computes the sequence of level values that maximizes the total probability along the path from the first to the last vowel of a sentence.

After the training is completed, this model was applied to the segmented database of the Cathy voice: For each prompt the frontend computes the Lesseme sequence and the supra-segmental features. Some of these features and the normalized acoustical parameters are then processed by the complete network and the Viterbi algorithm to compute the levels, which then replace the level information computed by the frontend. The approach was in part validated by inspection of the results, namely simply by comparing the level values computed by the frontend with those of the rewrite algorithm. It appeared that the choices of the rewrite algorithm were similar to what could be expected from listening to the prompts.

5. Results

In Blizzard Challenge 2013 we participated in two tasks, EH1 and EH2. The label for our system was the letter N. Our new EH2 system ranked much higher on perceptual listening tests than our old EH1 system. For the EH1 task we used the same synthesis method as we had in Blizzard Challenge 2012, in this case using a voice database for Cathy that we had already built earlier. While in Blizzard Challenge 2012, our system ranked in 2nd place (with the Greenman voice from Librevox), our 2013 EH1 system with the Cathy voice built using the same methods as our 2012 system fell to roughly rank 7 in 2013.

Given our approach to the EH1 task, there were effectively two benchmarks which allow a rough comparison between the 2012 and 2013 results. The overall median raw scores of both our EH1 voice, and the Festival benchmark voice, system B, each fell by 10 points from 2012 to 2013, from 31 to 21 for our EH1 Voice, and from 27 to 17 for the Festival benchmark. Thus, it appears that perceptual listening median raw scores are not invariant, but rather dependant on the overall quality of the systems entered. Higher median raw scores become more difficult to achieve as the overall quality of the systems being compared improve. Interestingly, this does not seem to apply to natural speech. In 2012, the original Greenman recordings were rated 46 by listeners, and in 2013 the Cathy recordings were rated 49; both are close to a perfect 50; the slight difference is probably attributable to the better overall quality of the professional Cathy recordings

For the second task, EH2, where we extended the voice database and used automatic relabeling to adjust prominence levels (see previous section) our results were better. Depending upon the specific attribute being ranked, there appears to be a group of 3 to 7 systems that usually have the highest scores. Given that our system was developed with prosody as one of its core elements, we tend to rank slightly higher in the paragraph versus sentence listening comparisons. For the overall paragraph ranking, our system is fourth in a group of seven systems that are pairwise insignificantly different than the highest ranked system L. In the following this is investigated in more detail for the partial tasks of EH2, using the data provided by the statistical data from the pair-wise Wilcoxon signed rank tests. We restrict ourselves to the statistics labeled with "all listeners". In any of the subsequent ranking attempts it needs to be kept in mind that overall and in each subtask there are fewer ranks than participants because of ties. The system A, which is natural speech, was excluded.

The rank ordering in the evaluation task "overall impression", shows our system (N) ranked 4th according to the median score. However, the differences within the group of the first 7 systems with the highest scores, corresponding to the system letters ***LKMNCIF*** (our system N is underlined) are all mutually insignificant, which can be seen from the pairwise significance matrices.

For more specific attribute data, systems M and K always ranked above Lessac's system N. Lessac ranked above the other four high ranking systems L, C, I and F on one or more of the measured attributes, and ranked above system F on all attributes.

For the listening task "pleasantness" the rank ordering starts with ***MKNL,*** showing our system 3rd. The difference to the first two, M and K, is significant, but the difference to system L is insignificant.

For the "speech pauses" task we were 7th, whereby the difference with the next 4 higher or equal ranking systems is insignificant and the differences to the next two systems with lower scores is insignificant.

The listening task "stress" put our system 6th, whereby the difference with the 3rd through 5th and 7th through 9th scoring systems is insignificant.

In the task "intonation", our rank is 4th tied with three other systems. The system ranked 3rd is better but with insignificant difference, while all systems with lower score are significantly different.

For the task labeled "emotion", our system ranks 4th. The rank order sequence here starts with ***MKLNCI***, whereby the differences to the system L, as well as to systems C and I are insignificant.

For "difficulty of listening" our system ranks 6th, but with insignificant difference to the systems ranked 3rd, 4th and 5th. The difference between our system and all systems ranked lower was significant.

For the partial listening task for sentences for task EH2, which were either from news or from novels, the systems performed as follows.

The system ranking for "similarity to the original" was ***MKNLC.*** Our system **N** ranked 3rd. The difference between systems M, K, N, and L was insignificant. All systems ranked below C were significantly different.

The system ranking for "naturalness" was ***KMLN***. Our system N ranked 4th. The difference between K & M was significant, but the differences between M & L, and L & N were not significant, although the difference between M & N was significant. All systems ranked below our system N were significantly different.

For the semantically unpredictable sentences (SUS) for task EH2, we received a median word error rate of 28 percent and a mean rate of 33 percent. The Wilcoxon signed rank test showed no significant difference to most of the other systems except to system C with the lowest median word error rate of 7% and to the next 5 systems with higher median word error rates of 14%. Interestingly the word error rates for our and most other systems were much lower when listened to by "paid listeners" (14/22 percent median/mean) and higher when "speech experts" were listening (43/46).

6. Conclusions

Overall our system's performance was close to or part of the group of 2-4 leading systems. In the partial attribute tests, often the differences between the best 1 to 4 systems and our system were insignificant, but differences to most systems with lower scores were significant.

The voice material used in the experiments was not typical for our methods of voice building. This was also the case in the Blizzard 2012 challenge, where our system ranked second, but this time the disadvantages of using voice material that was not recorded according to the methods underlying our design were more manifest. As explained in section 4, we made some attempt to relabel units to make better use of the speech material without the need to come up with a different design for prosodic rules that could have been more appropriate for this speaker. This automatic relabeling was partially successful but will need further study and refinement. The method of relabeling only level information was in some way rather crude. In particular, the Markov state model could have been more detailed by modeling the system dynamics of several salient intonation features together and taking consonantal context of the vowels into account not only in the static features but also in the Markov model.

7. References

[1] Lessac, A., The Use and Training of the Human Voice: A Bio-Dynamic Approach to Vocal Life, McGraw-Hill, 1996.

[2] Nitisaroj, R. and Marple, G. A., "Use of Lessemes in text-to-speech synthesis", in M. Munro, S. Turner, A. Munro, and K. Campbell [Eds], Collective Writings on the Lessac Voice and Body Work: A Festschrift, Llumina Press, 2010.

[3] Jordan, M. I. and Jacobs, R. A., "Hierarchical Mixtures of Experts and the EM Algorithm", Neural Computation, 6:181-214, 1994.

[4] Ma, J., Xu, L. and Jordan, M. I., "Asymptotic Convergence Rate of the EM Algorithm for Gaussian Mixtures", Neural Computation, 12:2881-2900, 2000.

[5] Larochelle, H., Bengio, Y., Louradour, J., and Lamblin, P. "Exploring Strategies for Training Deep Neural Networks", Journal of Machine Learning Research 1 (2009) 1-40